Safe Breast Cancer Diagnosis Resilient to Mammographic Adversarial Samples

Degan Hao[1], Dooman Arefan[2], Margarita L. Zuley[2], Wendie A. Berg[2], and Shandong Wu[1,2,3,4(✉)]

[1] Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA
[2] Department of Radiology, University of Pittsburgh, Pittsburgh, PA, USA
[3] Department of Bioengineering, University of Pittsburgh, Pittsburgh, PA, USA
[4] Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA
wus3@upmc.edu

Abstract. Adversarial data can lead to malfunction of deep learning applications. It is essential to develop deep learning models that are resilient to adversarial data while accurate on standard, clean data. In this study, we focus on building safe breast cancer diagnosis models against mammographic adversarial samples. We proposed a novel adversarially robust feature learning (ARFL) method to facilitate adversarial training using both standard data and adversarial data, where a feature correlation measure is incorporated as an objective function to encourage learning of robust features and restrain spurious features. To show the efficacy of ARFL for robust breast cancer diagnosis, we built and evaluated deep learning diagnosis models using two independent clinically collected breast imaging datasets, comprising a total of 9,548 mammogram images. We performed extensive experiments showing that the ARFL method outperformed several state-of-the-art methods. ARFL can serve as an effective method to enhance adversarial training, towards building safe breast cancer diagnosis against adversarial attacks in clinical settings. The code repository of this study is publicly available at GitHub: https://github.com/usernamesafeai/ARFL.

Keywords: Breast cancer diagnosis · Adversarial defense · Mammogram · Safe AI

1 Introduction

Adversarial samples can fool a deep learning classification model, where small and intentional perturbations may lead to unexpected results [24]. Adversarial attacking methods, such as projected gradient descent (PGD) [14], have shown success on attacking classification of natural view images. Adversarial attacks also pose threats to deep learning-based medical applications, such as inducing unsafe diagnosis, fraudulent insurance claims, biased clinical trial outcomes, etc. [6]. In the medical imaging domain, previous studies showed adversarial samples may downgrade a model's performance, as observed in image classification,

R. Mann et al. (Eds.): Deep Breath 2024, LNCS 15451, pp. 232–243, 2025.
https://doi.org/10.1007/978-3-031-77789-9_23

detection, and segmentation [16,18]. It is critical to develop deep learning models that are resistant to adversarial samples/attacks in order to deliver safe artificial intelligence (AI)-enabled medical applications.

Adversarial training, which trains a model by using a set of adversarially generated samples, is one of the few approaches to defend adversarial attacks [20]. Studies showed that by using the minimax optimization, adversarial training can improve a model's adversarial robustness [14]. Adversarial samples may also serve as a special type of data augmentation to increase a model's performance on the standard data (i.e., original clean data without adversarial perturbations) [26]. In the medical imaging domain, adversarial training-based methods have shown improved image diagnosis performance on either standard data [8] or adversarial data [10]. However, it remains challenging for a model to maintain stable performance simultaneously on both the standard data and adversarial data [9,17,19,25,28]. A previous study [12] indicated that the lack of exploiting the underlying manifold of data may be a key reason for this challenge.

While adversarial training has the benefits of resisting adversarial attacks, previous theoretical studies [19,25] showed that adversarial training at the same time may lower a model's performance on standard data, which is undesirable, as it is equally important to maintain the model performance on both standard data and adversarial data [17,28]. A recent study showed that adversarial training could result in even worse results when training with limited data [3]. To ensure stable model performance on both standard and adversarial data, a common approach is to merge the datasets for training [24], though this may fail when their distributions significantly differ. Researchers have considered standard data and adversarial data as two different domains to learn domain-invariant representations [22]. Another approach, as proposed in a recent work [1], is to perform training with separated batch normalization layers for standard data and adversarial data. Since the testing data's distribution is usually unknown in priori, it is difficult for this approach to choose which batch normalization layer to use. Another method, TRADES [28], demonstrates there may be a theoretical trade-off of the performance between standard and adversarial data. Overall, it remains an open research question in developing effective training methods to reconcile model performance on standard data and adversarial data.

In this study, we proposed a novel regularization method to build a breast cancer diagnosis model that is adversarially robust on both standard data and adversarial data. Our approach incorporates a feature correlation measure as an objective function, promoting robust features and reducing spurious ones when training on a mix of standard and adversarial mammogram images. We name our method ARFL (Adversarially Robust Feature Learning). Implemented on two real-world mammogram datasets (9,548 images total), ARFL's performance was compared with and without its integration, as well as against domain-specifi batch normalization (DSBN) method [1], TRADES [28], and multi-instance robust self-training (MIRST) [23]. Extensive experiment results on the two datasets showed the clear benefits of ARFL in maintaining the model's performance on both the standard data and adversarial data, and that our method outperformed the compared methods.

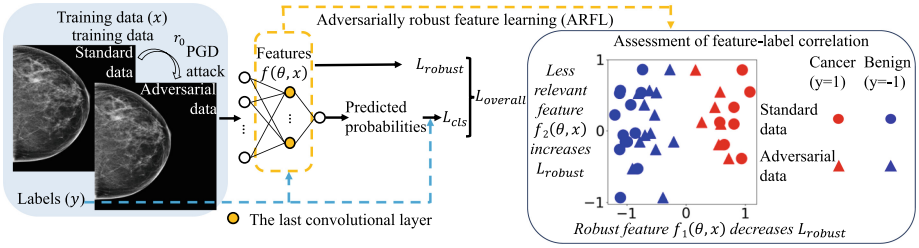

Fig. 1. Overview of the Adversarially Robust Feature Learning (ARFL) framework for breast cancer diagnosis. This figure shows the ARFL architecture using both standard and adversarial mammographic data as inputs. The adversarial training with ARFL focuses on extracting robust features $\mathbf{f}(\theta, \mathbf{x})$ for computing the robust loss $\mathbf{L_{robust}}$.

2 Related Work

AI has shown promise and early success in enhancing various tasks for medical image analysis, including detection, classification, segmentation, reconstruction, registration, etc. [2]. AI-based breast cancer diagnosis models are under active development and clinical translation [13]. It is imperative to ensure the deployment of such AI models are safe to patients, secure to clinical environments, and resilient to adversarial samples/attacks.

Adversarial security of AI models has attracted attention in the medical domain [8,15,27]. Such studies on breast cancer/imaging is scarce, but more challenging, as malignancy information in breast imaging may be more subtle and heterogeneous [11]. Researchers showed that adversarial mammogram images produced by generative adversarial networks can fool both breast cancer diagnosis models and experienced radiologists [29]. MIRST was introduced to defend adversarial attacks on breast ultrasound images [23].

3 Methods

3.1 Adversarially Robust Feature Learning (ARFL)

When training a classification model with both standard and adversarial data, the model simultaneously fits two potentially different distributions. As shown in Fig. 1, to encourage the learning of useful features from the mixed input and to reduce the chances the model learns from spurious correlations between the training data and truth labels, we introduced a regularization term, called adversarially robust feature learning (thus the name ARFL). As pointed by a previous work [9], a feature's usefulness can be measured by the expectation of feature-label multiplication, i.e., $\mathbb{E}_{(x,y)\sim\mathcal{D}}(f_{i,j}(\theta, x) \cdot y))$, and the feature is called ρ-useful if the expectation is greater than ρ. Inspired by such a correlation measurement, we designed a new loss function, named robust loss (denoted by L_{robust}), to characterize the feature-label correlation. L_{robust} is calculated by

summing up the absolute values of the product of each feature and label over the feature map, as shown in Eq. 1.

$$L_{\text{robust}}(\theta, x, y) = -\frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \sigma(\text{abs}(f_{i,j}(\theta, x) \cdot y)) \tag{1}$$

where input x can be either a standard input with an underlying distribution of $\mathbb{D}$; or an adversarial input from distribution $\mathbb{D}'$; H and W respectively denote the width and height of a feature map of input x; $\text{abs}(\cdot)$ denotes the absolute value function; $\sigma(\cdot)$ denotes the sigmoid function that scales the feature-label correlation; y denotes a positive or negative label $\{\pm 1\}$; θ denotes the model parameters; $f_{i,j}(\theta, x)$ denotes the value of the feature map at position (i, j). Considering that features near the output of a classification model contain more high-level information, we obtain the feature map from the last convolutional layer. L_{robust} encourages the model to learn features that are highly correlated with the labels. Different from the original method in [9], we revised the method to measure useful features by adding an absolute-value operation to consider both positive and negative correlations, and we also incorporated a sigmoid function to squash extreme loss values. Our method is appropriate as features showing either low positive correlations (yielding $\rho > f_{i,j}(\theta, x) \cdot y > 0$) or low negative correlations (yielding $-\rho < f_{i,j}(\theta, x) \cdot y < 0$) tend to be potentially less robust, leading to higher L_{robust} values. Then we integrate the adversarial loss and the robust loss as an overall loss for standard data as expressed in Eq. 2.

$$L_{\text{overall}}(\theta, x, y) = L_{\text{cls}}(\theta, x, y) + \lambda \cdot L_{\text{robust}}(\theta, x, y) \tag{2}$$

where L_{cls} denotes the binary cross entropy loss for binary classification tasks and λ is a weighting factor controlling the two objectives, i.e., the cross-entropy loss L_{cls} and the robust loss L_{robust}.

3.2 Integrating ARFL Into Minimax Optimization

To construct adversarial data, we introduced some degree of adversarial perturbation generated by PGD [14] to standard data (x). PGD generates adversarial perturbations by iteratively maximizing the perturbation towards the direction of changing the predicted output. To defend the adversarial attacks, adversarial training minimizes the loss of fitting the adversarial data while maximizing the same loss for the generated adversarial samples, as shown in Eq. 3.

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta \in \Delta(X)} L_{\text{cls}}(\theta, x + \delta, y) \right] \tag{3}$$

where δ denotes the perturbation imposed to x within the specified set of valid perturbations Δ, and y denotes the truth label.

With both standard data and adversarial data in each training batch, we minimize the empirical loss by fitting both the standard data and adversarial

data. We introduce Eq. 4 to implement the minimax optimization process.

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \Big[(1-r) \cdot \max_{\delta \in \Delta(X)} L_{\text{cls}}(\theta, x+\delta, y) + r \cdot L_{\text{cls}}(\theta, x, y) \Big] \tag{4}$$

where r denotes the ratio of the amount of standard data relative to the total amount of the data (standard data plus adversarial data) in each training batch. After integrating ARFL into adversarial training, we propose Eq. 5 for the minimax optimization on both standard and adversarial data.

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \Big[(1-r) \big(\max_{\delta \in \Delta(X)} L_{\text{cls}}(\theta, x+\delta, y) + \lambda \cdot L_{\text{robust}}(\theta, x+\delta, y) \big) + r \cdot L_{\text{overall}}(\theta, x, y) \Big] \tag{5}$$

where r can take various values in the range $[0, 1]$ to define different training schemes. The term L_{overall} is defined in Eq. 2.

4 Experiments and Results

4.1 Datasets

Our study was approved by the Institutional Review Board. We examined the effects of our method on two real-world mammogram imaging datasets for breast cancer diagnosis. The first dataset is from University of Pittsburgh Medical Center (UPMC) and the second is the publicly available Chinese Mammography Database (CMMD) [4]. The UPMC dataset was collected from a cohort of 1,284 women who underwent full field digital mammography screening. Each patient had one digital mammogram exam with up to four images of the two breasts (left craniocaudal [CC] view, left mediolateral oblique [MLO] view, right CC view, and right MLO view). Based on biopsy results, there are 366 patients diagnosed with breast cancer and 918 benign/negative cases. There are a total of 4,346 images. The images were acquired by a Hologic Lorad Selenia mammography system. The UPMC dataset is an internal private dataset and may be available to interested users upon request, after an approval from the institution along with a signed data use agreement and/or a material transfer agreement. The CMMD dataset was collected from a cohort of 1,775 patients who underwent mammography examination with both CC and MLO views. Based on biopsy, 1,310 patients are diagnosed with breast cancer and 465 patients are benign/negative, and there are a total of 5,202 images. The images were acquired by a GE Senographe DS mammography system. Using the two independent datasets, our target task is to perform computer-aided diagnosis of classifying breast cancer (i.e., malignancy) vs. benign/negative findings at patient level. The CMMD dataset is publicly available and can be downloaded from https://www.cancerimagingarchive.net/collection/cmmd/.

4.2 Experiment Settings

Model Structure and Training Settings: We used the VGG16 model [21] pre-trained on ImageNet [5] as the backbone. We fine-tuned the fully connected and last convolutional layers for binary classification of breast cancer. We implemented three training settings with parameter r: 1) standard training ($r = 1$), 2) adversarial training ($r = 0$) [14], and 3) dual adversarial training ($r = 0.5$) [10]. We trained with and without ARFL, setting L_{robust}'s weight λ to 10.0. Each model was trained for 100 epochs on both datasets.

Adversarial Sample Generation: We used PGD for adversarial attacks, with 7 iterative steps and an adversarial perturbation budget ε_1 of 0.01. The attacking perturbation budget ε_2 was set to 1e−4 to be visually imperceptible.

Comparison with Related Methods: We compared our method to three related methods, including DSBN [1], TRADES [28], and MIRST [23]. DSBN is a domain adaptation technique that allocates domain-specific affine parameters for data from different domains. DSBN was tested for adversarial training with standard data and adversarial data perturbed by the FGSM algorithm [8]. We replaced FGSM [7] with PGD [14], aiming to measure our method's resilience against this more threatening challenge. TRADES is an adversarial defense method that balances model performance on adversarial data and standard data using KL-divergence for regularization. MIRST uses different levels of perturbations to generate adversarial examples as additional data for self-training.

Performance Metric and Statistical Significance: We evaluated performance using the Area Under the Curve (AUC) and the standard deviation under five-fold cross-validation, where at each fold, 70% of the data for training, 10% for validation, and 20% for testing. Statistical significance was determined using the Mann-Whitney U test.

Visual Assessment: To visually assess feature learning effects using ARFL, we plotted feature saliency maps of mammogram images, calculated as gradients of loss with respect to the input.

4.3 Robustness Analyses of Hyperparameters

We analyzed the effects of the standard data mixing ratio (r), the weighting factor (λ), and the adversarial perturbation budget (ε_1) on model performance.

Effects of Mixing Ratio (r). We examined the effects of mixing standard data with adversarial data at varying ratios (i.e., robustness analysis of parameter r in Eq. 5). While in dual adversarial training where r is set to 0.5, it is interesting to examine whether other values of this ratio may lead to different performance. In this experiment, we measured the diagnosis model's performance additionally at $r = 0.25$ and $r = 0.75$ and compared to the effects when $r = 0.5$.

Effects of Weighting Factor (λ). The weighting factor λ, which controls the influence of L_{cls} and L_{robust} in the model, was varied from 0.1 to 100.0.

We applied ARFL in the context of dual adversarial training to determine the optimal balance point, where the model efficiently learns robust features without compromising classification performance.

Effects of Adversarial Perturbation Budget (ε_1). We investigated the impact of varying the adversarial perturbation budget ε_1 within the range of 0.005 to 0.1. We used 0.1 as the upper bound considering literatures and characteristics of mammogram images. Using the PGD method, we generated adversarial data constrained by this budget and incorporated the data into the adversarial training process. The aim was to observe how different levels of adversarial perturbation during adversarial training influence the model's defense against adversarial attacks.

5 Results

Table 1 and Table 2 show the mean AUC values and standard deviations on the test set of standard data and the test set of adversarial data, when using the UPMC dataset and CMMD dataset, respectively. As can be seen in Table 1, adversarial test had a substantially dropped performance under standard training (row A), which is the expected behavior for a standard model when facing adversarial attacks. When the model is trained by adversarial training (row C), adversarial test performance increased but at the same time the model downgraded in standard test - this sacrifice is undesirable for the slight benefit of adversarial robustness. When using dual adversarial training (row F), model performance largely increased in both standard test and adversarial test, showing the efficacy of this training method.

Table 1. Model performance comparisons on the UPMC dataset.

Training Method	Standard AUC	Adversarial AUC
A. Standard training	69.2 (1.1)	58.8 (1.4)
B. Standard training + ARFL	70.0 (1.9)	58.3 (3.5)
C. Adversarial training	61.7 (4.0)	56.9 (5.3)
D. Adversarial training + ARFL	62.5 (4.3)	59.2 (4.0)
E. Dual adversarial training	65.7 (5.9)	59.6 (9.4)
F. Dual adversarial training + ARFL	69.3 (2.3)	67.8 (2.4)
G. DSBN [1]	54.1 (8.5)	54.7 (9.0)
H. TRADES [28]	63.7 (3.5)	63.2 (3.5)
I. MIRST [23]	63.0 (1.9)	63.6 (1.7)

In terms of the benefits of ARFL, as shown in rows B, D, and F, while ARFL did not make a change in standard training (this is expected as ARFL is designed to mainly account for the mix of standard and adversarial data),

it largely improved the performance for adversarial training (row D) and dual adversarial training (row F; here the benefits are the highest), showing the usefulness of our proposed method, in not only resisting adversarial attacks but also maintaining the performance in the original standard data. In the comparison, DSBN (row G), TRADES (row H), and MIRST (row I) exhibited lower performance compared to dual adversarial training with ARFL (row F). The underperformance of DSBN can be attributed to its limitation in selecting specific batch normalizations for test sets. Furthermore, this comparison highlights that ARFL's approach of regularizing through feature-label correlation is more robust than TRADES, which regularizes with prediction-label correlation. It also demonstrates ARFL can learn robust features without using multiple instances as MIRST does.

Table 2. Model performance comparisons on the CMMD dataset.

Training Method	Standard AUC	Adversarial AUC
A. Standard training	64.9 (4.2)	41.5 (3.7)
B. Standard training + ARFL	64.9 (4.4)	41.5 (4.2)
C. Adversarial training	45.5 (4.6)	43.7 (4.7)
D. Adversarial training + ARFL	48.6 (4.5)	45.7 (4.7)
E. Dual adversarial training	67.8 (3.3)	66.3 (3.3)
F. Dual adversarial training + ARFL	68.8 (3.3)	67.3 (3.4)
G. DSBN [1]	54.7 (6.9)	55.5 (2.7)
H. TRADES [28]	64.8 (5.0)	61.9 (5.1)
I. MIRST [23]	64.4 (2.6)	64.8 (2.8)

When examining the results of CMMD shown in Table 2, a very similar overall performance pattern is observed as seen in Table 2, which further verifies the efficacy and generalizability of our proposed method on an independent dataset. The dual adversarial training with ARFL also outperformed DSBN, TRADES, and MIRST. In addition, on both datasets, the AUCs of the dual adversarial training with ARFL are significantly higher (all $p < 0.05$) than the AUCs of the adversarial training with ARFL.

It is worth mentioning that in Table 2 we noticed the adversarial training (row C) did not improve adversarial AUC compared to standard training (row A), though the standard deviation of the AUCs is also larger in row C compared to row A, showing the data heterogeneity may be higher in the CMMD dataset and that may lead to what we observed. Also note that the improvement resulted

from adversarial training is also modest under adversarial test on the UPMC dataset (Table 1, row C vs. row A). Previous studies showed that adversarial training may only improve adversarial AUCs under the use of a very large dataset [3]. This may partly explain the slight improvement observed in our study as our data scale is relatively small compared to large datasets.

Figure 2 illustrates on example mammogram images and the feature saliency maps for models trained with dual adversarial training with and without ARFL. In these maps, regions with sharp intensity contrast indicate important features, where higher gradients suggest stronger influence on the classification performance [25]. The comparison shows that incorporating ARFL results in a greater number of sharply contrasted regions, suggesting that ARFL enhances the learning of discriminative imaging features for the diagnosis purposes. Note that we demonstrate the saliency maps mainly on standard data as these clean data are better cases to illustrate and perceive the effects.

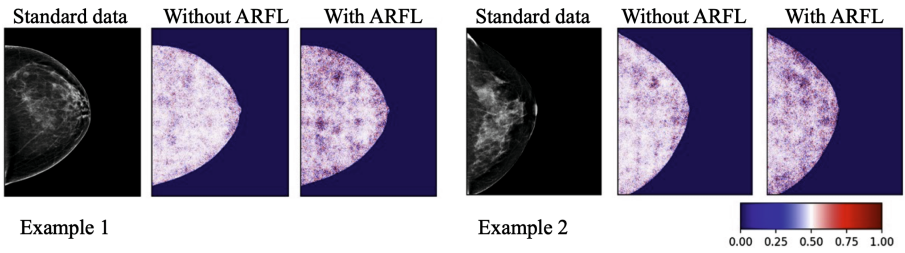

Fig. 2. Feature saliency maps of mammogram images from models trained using dual adversarial training with and without the integration of ARFL. The color bar represents the scaled gradients between zero and one. More regions with sharp contrast indicate more important features. (Color figure online)

Figures 3 shows the robustness analysis results. The sub figures in the left column shows model performance for varying r. In the UPMC dataset, $r = 0.5$ achieved the highest performance, while in the CMMD dataset, $r = 0.75$ was optimal. For consistency, results with $r = 0.5$ were reported to fairly compare with previous studies [10,24]. The sub figures in the middle column shows the effects of adjusting λ. The highest test AUC was achieved at $\lambda = 10.0$. The right sub figure shows the model's test AUCs for varying ε_1. As ε_1 increased, AUC initially increased, then stabilized at 0.01 and beyond. This suggests an optimal range for ε_1 in adversarial training for our study/data. These experiments supported the use of optimal parameter values in our main experiments. Note that optimal values may differ for other datasets or tasks.

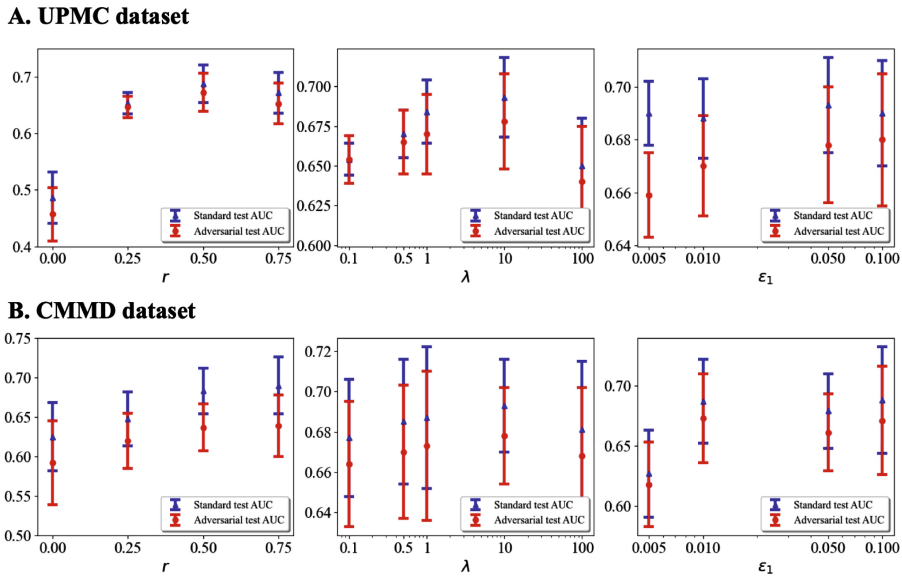

Fig. 3. Robustness analysis of hyperparameters: standard data mixing ratio (r), weighting factor (λ), and adversarial perturbation budget (ε_1). Shown are AUC values with varying values of hyperparameters. Error bars represent standard deviations.

6 Conclusion

In this work, we designed a novel method, ARFL, to facilitate adversarially robust adversarial training for safe breast cancer diagnosis. ARFL facilitates the learning process towards identifying features that are strongly correlated with true labels. On the two breast mammogram datasets, ARFL showed benefits in resisting adversarial samples and maintaining stable diagnosis performance on standard data. Our extensive experiments on the two datasets from different sources showed similar efficacy and the generalizability of our method. ARFL also outperformed the compared methods. For future work, we will extend the evaluation of our method on other imaging data and other types of adversarial attacks.

Acknowledgement. This work was supported by the 1R01EB032896 grant (and a Supplement grant 3R01EB032896-03S1) as part of the NSF/NIH Smart Health and Biomedical Research in the Era of Artificial Intelligence and Advanced Data Science Program, a NSF grant (CICI: SIVD: #2115082), the Jewish Healthcare Foundation RAPS Seed Grant Program, and the University of Pittsburgh Momentum Funds for the Pittsburgh Center for AI Innovation in Medical Imaging. This work used Bridges-2 at Pittsburgh Supercomputing Center through allocation [MED200006] from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by NSF grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

1. Chang, W.G., You, T., Seo, S., Kwak, S., Han, B.: Domain-specific batch normalization for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7354–7362 (2019)
2. Chen, X., et al.: Recent advances and clinical applications of deep learning in medical image analysis. Med. Image Anal. **79**, 102444 (2022)
3. Clarysse, J., Hörmann, J., Yang, F.: Why adversarial training can hurt robust accuracy. arXiv preprint arXiv:2203.02006 (2022)
4. Cui, C., et al.: The Chinese Mammography Database (CMMD): an online mammography database with biopsy confirmed types for machine diagnosis of breast. The Cancer Imaging Archive (2021)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
6. Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S.: Adversarial attacks on medical machine learning. Science **363**(6433), 1287–1289 (2019)
7. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
8. Han, T., et al.: Advancing diagnostic performance and clinical usability of neural networks via adversarial training and dual batch normalization. Nat. Commun. **12**(1), 1–11 (2021)
9. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
10. Joel, M.Z., et al.: Using adversarial images to assess the robustness of deep learning models trained on diagnostic images in oncology. JCO Clin. Cancer Inform. **6**(2), e2100170 (2022)
11. Kim, J.H., et al.: Breast cancer heterogeneity: MR imaging texture analysis and survival outcomes. Radiology **282**(3), 665–675 (2017)
12. Lin, W.A., Lau, C.P., Levine, A., Chellappa, R., Feizi, S.: Dual manifold adversarial robustness: defense against LP and non-LP adversarial attacks. In: Advances in Neural Information Processing Systems, vol. 33, pp. 3487–3498 (2020)
13. Lotter, W., et al.: Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. Nat. Med. **27**(2), 244–249 (2021)
14. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
15. Mirsky, Y., Mahler, T., Shelef, I., Elovici, Y.: CT-GAN: malicious tampering of 3D medical imagery using deep learning. In: 28th USENIX Security Symposium (USENIX Security 19), pp. 461–478 (2019)
16. Paschali, M., Conjeti, S., Navarro, F., Navab, N.: Generalizability vs. robustness: exploring adversarial examples in medical imaging. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2018. Springer (2018)
17. Picot, M., Messina, F., Boudiaf, M., Labeau, F., Ayed, I.B., Piantanida, P.: Adversarial robustness via fisher-rao regularization. IEEE Trans. Pattern Anal. Mach. Intell. (2022)
18. Qi, G., Gong, L., Song, Y., Ma, K., Zheng, Y.: Stabilized medical image attacks. arXiv preprint arXiv:2103.05232 (2021)

19. Raghunathan, A., Xie, S.M., Yang, F., Duchi, J.C., Liang, P.: Adversarial training can hurt generalization. arXiv preprint arXiv:1906.06032 (2019)
20. Shafahi, A., et al.: Adversarial training for free! In: Advances in Neural Information Processing Systems, vol. 32 (2019)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
22. Song, C., He, K., Wang, L., Hopcroft, J.E.: Improving the generalization of adversarial training with domain adaptation. In: International Conference on Machine Learning, pp. 4934–4943. PMLR (2018)
23. Sun, S., Xian, M., Vakanski, A., Ghanem, N.: MIRST-DM: multi-instance RST with drop-max layer for robust classification of breast cancer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 401–410. Springer (2022)
24. Szegedy, C., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
25. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A.: Robustness may be at odds with accuracy. arXiv preprint arXiv:1805.12152 (2018)
26. Xie, C., Tan, M., Gong, B., Wang, J., Yuille, A.L., Le, Q.V.: Adversarial examples improve image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 819–828 (2020)
27. Yao, Q., He, Z., Han, H., Zhou, S.K.: Miss the point: targeted adversarial attack on multiple landmark detection. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12264, pp. 692–702. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_67
28. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.I.: Theoretically principled trade-off between robustness and accuracy. In: International Conference on Machine Learning, pp. 7472–7482. PMLR (2019)
29. Zhou, Q., et al.: A machine and human reader study on AI diagnosis model safety under attacks of adversarial images. Nat. Commun. **12**(1) (2021)